R14/Ishank Juneja/16D070012

March 15, 2020

A well known obstacle to partially observable Markov decision process (POMDP) planning is the *curse of history*. The term refers to the problem that the optimal action in the current time step depends on the agents entire history which is exponential in the time horizon. To tackle the curse of history, the authors propose a new approximate POMDP planning approach called *Point-Based Value Iteration* (PBVI).

POMDP planning and Exhaustive Enumeration
A POMDP can be fully described by the tuple $(S, A, O, b_o, T, \Omega, R, \gamma)$, where S is a finite set of states, A is a set of discrete actions and O is the finite set of observations an agent can receive. T provides us with the transition function $T(s, a, s')$, b_o is the initial belief state and $\Omega(o, s, a)$ is the distribution describing the probability of observing o given the action a taken by the agent and the next state s reached by the agent. $R(s, a)$ is the usual reward signal and γ is the discount factor. Since the state is only partially observable, at every time step t the agent computes a belief state vector $b_t(s')$.
The goal of POMDP planning then is to take the optimal action a for every belief state b_t that it encounters during its lifetime. The POMDP planning problem can be solved using exact value iteration (VI) by exhaustively enumerating all possible value functions at every time step. In particular, we use the notation $V_n = \{\alpha_0, \alpha_1, \ldots, \alpha_m\}$ to denote the solution set for the optimal value function after n iterations. Here, in the terminology of the previous reading [R12 - Kaelbling et al.] each α-vector represents a certain policy tree in the policy space and the optimal value function $V_n(b)$ is given by $V_n(b) = \max_{\alpha \in V_n} \alpha \cdot b$. The simplest procedure to obtain the set V_n is then to build a superset of potential α values by building upon the (assumed to be minimal) set V_{n-1}. Subsequently this superset is pruned to remove any values that may lie entirely below a combination of other α vectors.
The problem with this approach is that in the worst case, the run time is orderwise equal to $|S|^2|A||V_{n-1}|^{|O|}$ which is exponential in the size of the observation space. This is quite undesirable, the authors use this drawback as their motivation for PBVI - an approximate solution method.

Point-Based Value Iteration (PBVI)
The underlying philosophy for this method is that planning equally for all belief states within the belief simplex is unnecessary, instead the focus should be on getting away with planning for as few *representative* belief points as possible.
The proposed PBVI algorithm solves the POMDP by planning for a finite set of belief points $B = \{b_o, b_1, \ldots, b_q\}$. This way only the optimal α-vectors corresponding to elements in B need to be optimized for during value iteration. Solving the problem in such an approximate manner makes POMDP planning tractable even for problems with larger, more complex states-spaces, since the run time is now polynomial and given by $|S||A||V_{n-1}||O||B|$.
To deliver quick results initially and improve its accuracy over time, PBVI performs *belief set expansions* at the end of every outer iteration. The performance of PBVI is completely dependent on the choice of belief set B. Keeping this in mind, the authors propose and compare four strategies for belief set expansion,

1. Random (RA) - New belief points are sampled from a uniform distribution over the entire belief simplex.
2. Stochastic simulation (SS) with random action (SSRA) - In all the SS based strategies, a single step forward of the trajectory is stochastically simulated to produce new belief states $\{b_{a_0}, b_{a_1}, \ldots\}$ i.e. a new b for every action. Under the SSRA scheme one of these actions is randomly selected.
3. SS with greedy action (SSGA) - The action selected and simulated is the one which is greedy with respect to our most recent estimate of the value function. The new belief state b' obtained in this manner is appended to the set B.
4. SS with explorative action (SSEA) performs one step simulations corresponding to all possible actions to generate the belief states $\{b_{a_0}, b_{a_1}, \ldots\}$. Among these, it keeps the new belief state b_{a_i} that is farthest away from any point already in B.

PBVI error and experimental validation
Based on performance comparisons on various problem domains, the authors select SSEA as their belief set expansion method. Subsequently they prove a bound on the error created by the PBVI approximation - $||V_n^B - V_n^*||_\infty$ for this scheme. Overall their PBVI solution proves to be quite effective on the Tag, Hallway and Maze33 domains out performing the QMDP baseline on all three after a reasonable computation time. Due to exact solutions being intractable on all but the simplest of POMDPs, efficient and provably accurate approximate planning is crucial. This work by Pineau and others is a big leap in this regard.

A question I have is -

- Why does the QMDP approximation work well on numerous POMDP domains? It seemed to me like some strong regularity conditions on the POMDP would be needed to make this claim in general.